突发公共卫生事件中微博谣言的识别*

■ 石锴文　刘勘
中南财经政法大学信息与安全工程学院　武汉 430073

摘　要：[目的/意义]在"新冠"疫情这类突发公共卫生事件中，网络社交媒体上迅速产生大量关于疫情的言论，其中包含不少蓄意传播的谣言，不仅危害公众心理健康，而且会影响应对公共卫生事件的方案实施。识别突发公共卫生事件的谣言能够使民众正确面对危机，为社会安定、网络治理起到积极的维护作用。[方法/过程]首先对采集到的疫情期间已被证实的谣言进行深度分析，提取谣言文本的主要特征，包括上下文特征、话题类别特征、情感程度特征、关键词特征等；然后针对文本分类模型中的文本特征表达较为单一的问题，利用不同的模型对提取的谣言文本特征进行向量化，并对各类文本特征进行加强和融合。其中通过 TF-IDF 计算的词向量权重在捕获上下文特征的同时，能够加强词粒度的关键词特征信息。最后，使用 BiLSTM + DNN 模型对融合的特征向量进行分类判别。[结果/结论]实验结果表明，话题类别、情感程度等特征对谣言识别均有贡献，特别是经过强化后的词向量与其他特征融合后对识别准确率有明显提升，召回率、F1 值等指标均达到 90% 以上，效果超过其他的谣言识别模型，说明笔者所构建的方法能够很好地实现对突发公共卫生事件背景下的谣言识别。

关键词：公共卫生事件　谣言识别　微博　多特征融合
分类号：TP393
DOI：10.13266/j.issn.0252 – 3116.2021.13.009

1 引言

根据《突发公共卫生事件应急条例》[1]，突发公共卫生事件定义为突然发生、造成或者可能造成社会公众健康严重损害的重大传染病疫情、群体性不明原因疾病、重大食物和职业中毒以及其他严重影响公众健康的事件。突发公共卫生事件与公众生活息息相关，与之伴随的是在微博、微信等社交媒体中引发的大量网络谣言，这些谣言的传播给公众心理稳定以及政府治理造成较大的阻碍，对社会安定和民生保障构成巨大威胁。例如在新冠疫情初期，"只有 N95 口罩具有防疫功效"的网络谣言误导群众大量抢购囤积 N95 口罩，严重影响了公众对病毒的正常防范。因此，突发公共卫生事件中的谣言识别紧迫且重要。然而，由于此类谣言具有迷惑性大、情感性强、关注程度高等特殊性质，使其识别难度较大。

微博是我国言论传播最广泛的平台之一，在突发公共卫生事件中起着信息传播和舆情引导的重要作用，但也是网络谣言滋生和扩散的途径。谣言在微博上曝光度大，其传播比微信、论坛等网络媒体范围更广、迷惑性更大、影响程度更深，因此笔者将研究范围聚焦在微博平台上的谣言，其中如何从公共卫生事件的谣言中提取有效特征尤其是内容特征对谣言识别起到关键作用[2]。针对微博的特征提取，不同研究视角关注的特征均有差异，现在较多的研究聚集在内容特征和用户特征上。内容特征关注微博言论的上下文特征、语义特征、多媒体特征等；用户特征则关注用户的行为、影响力等特征。由于突发公共卫生事件的特殊性，笔者将重点关注微博文本的内容特征，在分析突发公共卫生事件下微博文本与平时文本差异的基础上，使用不同的文本特征描述方法，从不同的角度分析和提取微博谣言的文本特征，然后将各种特征进行融合，构建有效的突发公共卫生事件下的微博谣言检测模型。

* 本文系中央高校基本科研业务费交叉学科创新研究项目"大数据支持下网络谣言的智能消解机制研究"（项目编号：2722021EK016）研究成果之一。
作者简介：石锴文（ORCID：0000 – 0002 – 3563 – 982X），本科生；刘勘（OCRID：0000 – 0001 – 9339 – 7315），教授，博士，通讯作者，E-mail：liukan@zuel.edu.cn。
收稿日期：2020 – 12 – 21　修回日期：2021 – 04 – 13　本文起止页码：87 – 95　本文责任编辑：徐健

2 相关工作

关于谣言识别的研究,大多从两个角度展开:谣言特征提取和识别算法设计,多数研究关注如何从谣言数据中提取有效的特征,一部分则关注识别谣言的分类算法。

在谣言特征方面的研究中,最早来源于C. Castillo等人[3]评估Twitter上的新闻可信程度,所提取的68个特征涵盖消息基本特征、用户特征、话题特征以及传播特征。之后F. Yang等[4]使用微博谣言数据,提出两个新的特征:用户使用的客户端以及事件发生的地理位置;贺刚等[5]提出一系列新的特征,包括符号特征、链接特征、词频分布特征和时间差等,并与微博文本特征及用户特征结合;夏松等[6]通过一种新设计的抽词算法构建敏感词库,在微博内容特征、用户行为特征等基本特征中加入敏感词特征使谣言识别准确率有明显提升;李钢等[7]提出基于受众年龄的新型谣言传播的耦合社交网络,从受众的认知能力、匿名程度、权威性等基本特征,以及受众的从众心理、记忆效应、好友的影响作用等心理特征方面对受众进行画像,构建多维度函数实现基于受众画像的谣言传播模型。

谣言的时间序列特征也倍受关注。S. Kwon等[8]首次指出谣言事件传播过程中时间属性的重要性,通过研究时间、结构、语言3个方面的传播特性来确定谣言的特征,然后构建随机森林分类器;J. Ma等[9]在S. Kwon的基础上,进一步扩展了随时间变化的特征集合,利用简单的等长时间序列划分来观察谣言事件特征随时间的变化,这个时间序列的建模技术被应用于整合各种社交语境信息;王志宏等[10]为了更好地观察和表示谣言事件特征随时间的变化,引入模糊时间序列模型中的论域划分思想,将事件的时间跨度作为论域,提出了基于模糊聚类的事件时序数据动态划分算法,并在此基础上构建了随时间变化的时间特征集合;M. Kotteti等[11]提出了一个多时间序列数据分析模型来检测Twitter上的谣言,所提出的方法仅使用推文的时间特性来代替检查推文的内容,这使得计算复杂度大大降低,从而可以快速检测谣言。

在谣言识别算法方面,除了传统的逻辑回归、支持向量机、随机森林等机器学习技术,越来越多的研究人员使用深度学习模型进行谣言检测。J. Ma等[12]利用递归神经网络模型对上下文提取的优势分析某一话题前后帖子的联系,以减少无价值帖子对谣言鉴别结果的影响;L. Li等[13]提出了一种基于深度双向门控递归单元(D-Bi-GRU)的谣言检测方法,通过捕获微博流的前向和后向上下文特征,获取随微博事件群体响应信息;王星宇[14]构建了3个深层特征,加入到基本的浅层特征中,其中两个深层特征为Doc2vec构建的微博内容句子向量,以及Snownlp情感分析库计算得到的情感两极分化程度;刘勘等[15]将迁移学习应用到Twitter谣言检测中;G. Siva等[16]构建了虚假新闻的图特征,然后基于图的特征向量学习和标签扩展算法进行无指导学习;而F. Marra等[17]则利用对抗生成网络识别社交网络中的虚假图片;类似地,F. Qian等[18]将文本生成技术应用到虚假新闻识别中,对新发生的新闻利用生成器产生针对此新闻的评论来判断其真伪。

但是目前针对突发公共事件或社会热点事件中的谣言识别研究并不多。樊荣等[19]在微博平台上选取2016年"山东非法疫苗事件"以及"米脂三中伤人事件"相关谣言文本,构建基于用户舆情历史文本、谣言关注度、微博频率的R-CNN识别模型;曾子明等[20]使用LDA主题模型和随机森林算法对2016年雾霾谣言进行了检测;王林等[21]基于ELM、TAM模型以及生命周期理论,使用信息内容、发布日期、发布者认证类型等特征变量建立突发公共卫生事件舆情传播影响因素模型;李丽华等[22]选取2017年英国5起暴恐袭击事件为研究案例,在Twitter的舆情数据集上,对舆情传播主体、信息内容特征以及传播特征进行分析,研究相关谣言的传播特点及机制。

综合来看,在谣言识别的特征提取方面往往来自研究者的专门设计,且大部分谣言聚焦在常规情景下。而在识别算法方面,基于深度学习网络的谣言识别需要大量的训练数据。针对突发公共卫生事件这类特定领域的谣言识别,需要研究和提取话题内容、语言特征、情感极性等符合公共卫生事件背景要求的特有特征。同时,不同特征之间的融合也将对谣言识别起到较大作用。因此,笔者将设计和提取多个谣言文本的内容特征,并探讨其有效的融合方法,构建特定突发公共卫生事件下的谣言识别模型。

3 微博谣言的特征分析

3.1 谣言数据来源

笔者以突发新型冠状病毒肺炎疫情为例,从微博社区管理中心的微博辟谣官方账号中收集已经被证实为谣言的数据,时间段为疫情爆发以来谣言较为集中的前4个月,即2020年1月1日到4月30日内的有关疫情谣言的微博,共计730条。由于很多微博消息看

起来可信度不高,但是真正被证实和辟谣的信息并不多,这说明人工辟谣是一件非常困难和耗时的事情,这也导致可用的标注数据大大减少,限制了深度学习模型的使用,因此研究重点为谣言的特征提取和特征融合。作为对比并考虑到数据平衡的问题,笔者另外通过随机采集和人工筛选的方式得到了疫情期间被证实的非谣言微博数据 1 400 条,这些数据取自新华网、人民网、央视网等官方微博,以保证其真实有效。数据采集使用 selenium 模拟浏览器获取网页信息,通过 BeautifulSoup 解析网页内容,最后用正则表达式与 find 函数匹配所需字段。图 1 显示了采集数据的样例。

微博辟谣
2020-2-29 22:03 来自 微博 weibo.com
#微博辟谣# 网传"益阳确定一例，传播者传给爷爷，并造成790人接触"系谣言

@益阳市公安局
#微博辟谣# 【网传"益阳确定一例，传播者传给爷爷，并造成790人接触"系谣言】2月29日，部分微信群、朋友圈传言"今天凌晨三点益阳确定一例，传播者为武汉大学学生，传给了爷爷，已造成790人接触，已隔离672人，118人还未找到。"经核实，该信息为谣言，系相关疫情信息剪辑混合、移花接木 ... 展开全文

（a）网络谣言样例

央视网
2020-2-29 20:00 来自 微博 weibo.com 已编辑
【#南开大学发放流量补贴#，所有学生都有份】为保障学生疫情期线上学习、线上求职，@南开大学 决定向全校29819名学生（包括本科生和硕士、博士研究生）以现金形式向银行卡发放每人100元专项补助，用于补贴疫情防控期间网络流量。网友：又是别人家的大学

（b）非谣言样例

图 1　疫情期间的谣言和非谣言数据

3.2　疫情谣言文本分析

突发公共卫生事件与一般类型事件的网络舆情比较而言,除具备网络舆情突发性、直接性、互动性、即时性等一般特征外,还具有公众参与度高、负面倾向性强、民众恐慌加剧和极端用语多等独有的特征,其中有关微博的文本特征主要表现在以下几个方面:

(1)谣言词云。根据谣言文本绘制了如图 2 所示的词云图,图中 "感染""预防""口罩""治疗""开学" 5 个词语最为突出。当突发卫生事件爆发时,民众们最主要的关注点就落到疾病及民生问题上。有关人民生命安全、衣食保障、出行途径、开学复工的消息成为大家最愿意获取的内容。造谣者深谙民众所想,往往顺应民众心理发布相关谣言,用以带来较大的关注,如预防新冠肺炎、治疗新冠肺炎,病毒传播途径、公共交通出行等。而且通常具有一些固定的句式,例如"…是…,望周知""只有…才能…,切记切记"。这些言论常常危言耸听,利用特定背景下民众高关注度的特点,形成舆论导向性。



图 2　疫情谣言词云图

(2)词频特征。微博疫情谣言中的词语出现的频次如表 1 左半所示,其中呈现一定的规律。有关"极""非常"等程度词语、有关"重大""重要"的词语、有关"一定""千万"等词语出现频率较高,这类词语往往看起来更有紧迫感。有关突发公共卫生事件的标志性词,如"病毒防治""预防""医院""专家""酒精"等频繁出现。作为对比,表 1 右半部分显示了非谣言的高频词,其中"近日""回应""实施"等词语较为正式,夸张的程度副词几乎不出现。

表 1　疫情谣言和非谣言的高频词

谣言高频词				非谣言高频词			
词语	词频	词语	词频	词语	词频	词语	词频
不要	48	医院	60	说	138	近日	283
重……	44	消毒	34	称	76	发生	232
极	28	酒精	27	回应	42	视频	213
千万	26	药	19	有人	39	转发	212
一定	16	防治	17	实施	36	可怜	154
非常	11	专家	10	流感	17	救救	135
……	……	……	……	……	……	……	……

(3)关键词特征。为了进一步分析谣言文本的用词特征,使用 TF-IDF(Term Frequency-Inverse Document Frequency, TF-IDF)方法计算各个词在不同谣言中的权重,从而反映出这些词的关键性。对疫情谣言进行 TF-IDF 计算后,得到的关键词如表 2 左半所示,其中关于中国、美国、武汉、封城、医疗、医院、疫情、肺炎等关键词的谣言最为突出。在表 2 右半所示的非谣言的文本中,直播、新闻、发布、生活、健康、资讯等关键词较为突出,主要集中在日常的信息资讯。

(4)主题特征。笔者利用 LDA 主题模型提取微博谣言文本的主题特征,使用主题向量结构相似度最小化方法[23]确定主题个数为 7,最终提取出的主题向量如表 3 所示,其中主要包括疾病防治、灾害救助、政策解读、人物聚焦、民生保障、疫情动态、科普知识 7 大主

表 2 疫情谣言关键词

疫情谣言关键词				非疫情谣言关键词			
词语	权重	词语	权重	词语	权重	词语	权重
美国	18.22	肺炎	8.80	直播	25.98	事故	9.20
中国	13.85	医院	8.66	新闻	24.73	生活	8.34
医疗	13.43	日本	8.21	孩子	20.61	全球	5.92
疫情	10.45	俄罗斯	8.16	网友	14.41	资讯	5.34
允悲	10.16	封城	7.90	发布	10.90	健康	5.24
开学	9.71	消毒	7.82	国家	9.96	希望	4.17
武汉	9.44	……	……	关注	9.83	……	……

题。在这些主题中,疾病防治、灾害救助和民生保障等都是一直伴随各类公共卫生事件的主题。而如“钟南山:确诊病例中没有素食者”“李兰娟说谈恋爱可以预防新冠肺炎”等聚焦人物的谣言;如“XX 地马上也要封城”“政府将用飞机喷洒消杀药物”等对政策曲解的谣言;如“吹风机吹口罩可消毒”“空调开到 20℃病毒就会死亡”等看似科普的谣言,都与新冠这样的特定事件紧密联系,也是民众最关心的主题,谣言传播者正是利用这些主题迷惑大众。

表 3 疫情谣言主题词汇统计

主题	词 1	词 2	词 3	词 4	词 5
疾病防治	设施	抢救	医疗	病毒	聚集
灾害救助	药物	捐赠	支援	调用	无偿
政策解读	违法	公布	复工	调任	解封
人物聚焦	英雄	演讲	医生	行程	重症
民生保障	交通	小区	快递	拦截	消毒
疫情动态	近日	日本	美国	出院	名
科普知识	药品	发生	降低	视频	应该

(5)情感特征。谣言的情感往往比正常言论更加强烈和丰富,笔者使用知网情感分析用词语集对谣言语句的情感进行分析,使用谣言及非谣言样本各 730 条,得到谣言情感得分的分布图,如图 3 所示。情感得分在 0 到 1 之间,接近 1 表示正面情绪,接近 0 则为负面情绪。可以发现热点事件的情感具有明显的情绪化特点,集中在两端分布。且相比于正常言论中正面情绪绝对的主导地位,谣言负面情绪的占比增加显著,已接近正面情绪的水平。

4 模型构建

4.1 基本思路

笔者将设计和提取多个谣言文本的内容特征,探讨有效的融合方法,并构建基于特定突发公共卫生事件下的谣言识别模型。通过多组对比实验,观察以往方法和笔者提出的方法在微博谣言数据集上的效果,

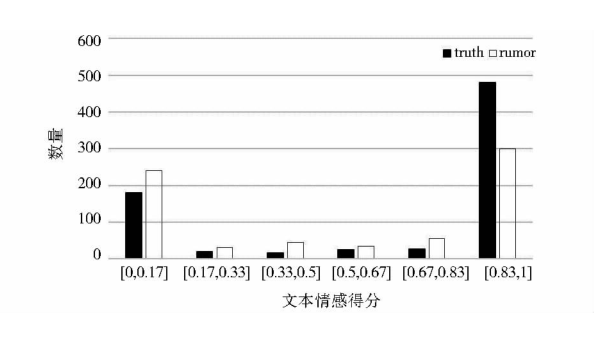


图 3 谣言情感得分的分布图

rumor 为谣言分布,truth 为非谣言分布

并对比不同特征组合后模型的分类性能。

针对上节分析的突发公共卫生事件下的微博谣言文本特征,笔者将重点利用主题类别特征、情感特征以及关键词特征来进行谣言识别,其中关键词权重计算时包含了词频特征。不同角度的特征可以相互补充,然后将各种特征向量化转换并进行特征融合,构建谣言的识别模型。融合过程包括对关键词特征进行文本增强,以及与主题特征、情感特征进行拼接,最后通过深度学习网络进行是否为谣言的类别判别。

4.2 模型过程

构建的谣言识别模型如图 4 所示,包含有基础特征提取层、特征融合层和分类判别层。基础特征提取层首先对输入的文本进行分词和去停用词的预处理,之后使用训练好的模型对输入的文本进行基础特征的计算,包括用 Word2Vec 模型计算的词向量语义特征、利用 TF-IDF 计算的关键词特征(f1)、利用 LDA 模型计算的主题特征(f2)和基于情感词典计算的情感特征(f3)。特征融合层包括两个部分:①利用关键词权重对词向量进行强化,突出关键词的权重;②将各类特征串联拼接在一起,得到最终的融合特征向量。分类判别层采用长短期记忆神经网络(Bi-directional Long Short-Term Memory, BiLSTM)和深度神经网络(Deep Neural Networks, DNN)对拼接后的特征向量进行分类训练,最终并输出类别标签。

(1)关键词增强。词向量特征是文本特征的重要内容,在整个识别模型中的重要一步也是对关键词向量的加权增强。在数据预训练时采用 Word2Vec 算法计算得到的词向量可以较好地提取文本的上下文特征,但此时的词向量重点不够突出,难以反映语句中的核心词语。将利用 TF-IDF 计算的权重特征加入,可以强化关键词的向量权重,使词向量重点突出,更好地体现文本特征。利用 TF-IDF 权重对词向量进行强化的流程见图 5。

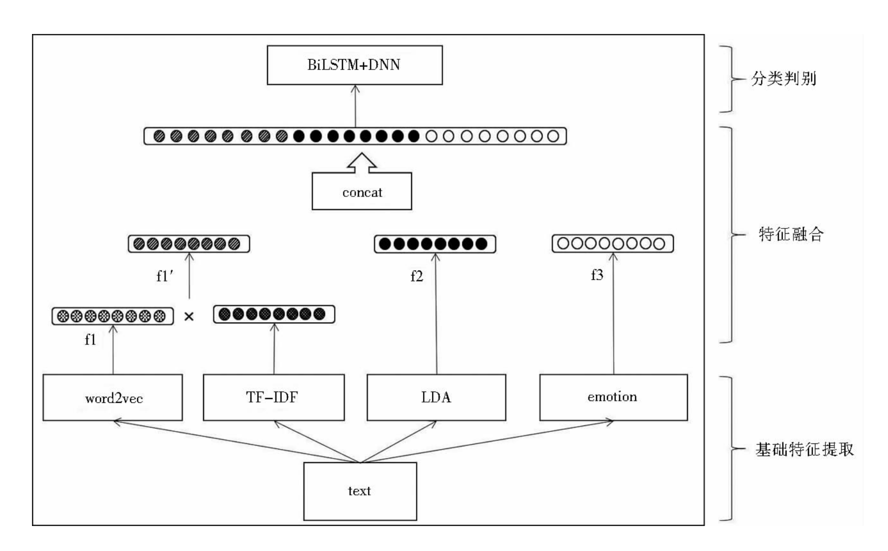


图 4　多特征融合的谣言识别模型

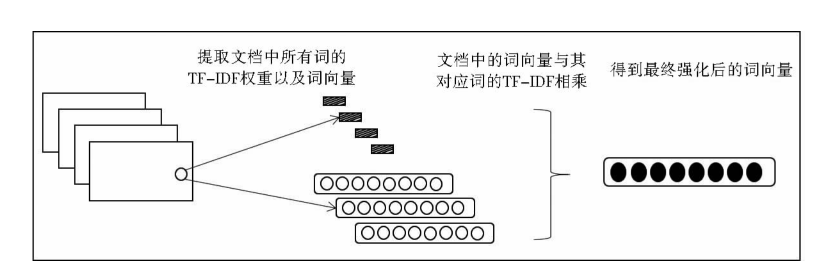


图 5　TF-IDF 强化词向量的过程

在图 5 中，设文档集为 D，用 d_i 表示其中的一篇经过分词和去停用词处理后的文档；用 W 表示该文档集的词表，用 w_i 表示词表中的一个词，公式(1)为文档集中的每一篇文档的向量化表示：

$$docVec(d_i)=\frac{C}{length(d_i)}\sum_{w_i\in d_i} tfidf_{d_i}(w_i)\cdot WordVec(w_i) \quad 公式(1)$$

其中，$length(d_i)$ 表示文档中词的数量，C 是一个避免梯度消失的权重调节系数，$tfidf_{d_i}(w_i)$ 表示相应文档中词的 TF-IDF 权重，$WordVec(w_i)$ 表示词 w_i 在整个语料 D 中的向量。上述公式本质是将对应文档中词的 TF-IDF 权重广播到该词的词向量中，使得词向量包含上下文以及关键词信息。通过 TF-IDF 权重的广播，相同词在不同的文档中则有了不同的向量表达，使得词的嵌入式表达含义更加丰富完整。

（2）特征拼接。在融合了 Word2Vec 与训练模型和 TF-IDF 关键词权重模型以后，得到了谣言中每个词新的词向量 f1′。利用 LDA 模型可以计算每条谣言在 7 个主题的概率分布，这样得到谣言的主题分布向量，这是一个 7 维向量。情感计算则利用 Snownlp 计算所得的得分进行归一化。最后将包含了这些特征的向量(f1′＋f2＋f3)拼接到一起，送入接下来的分类训练模型。

（3）谣言判别模型。特征融合之后就可以构建分类器，实现微博谣言的识别。笔者使用基于 BiLSTM＋DNN 网络的分类器模型，经过 BiLSTM 层进行特征抽取，将正反两个方向的序列化输出进行拼接，再送入 DNN 层与输出层，对样本是否为谣言进行预测。其具体结构如图 6 所示：

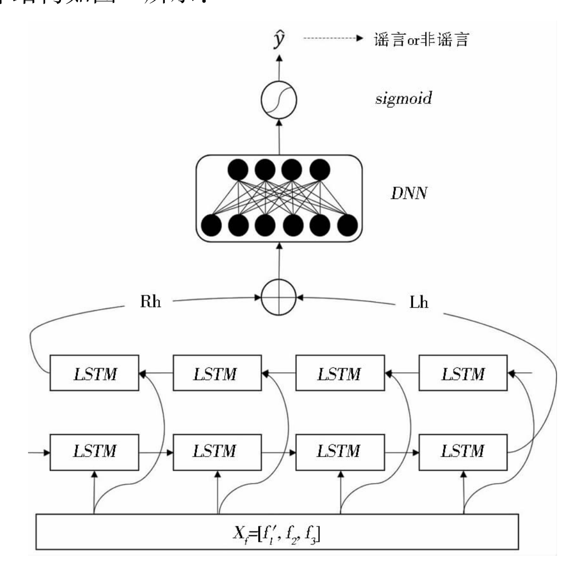


图 6　微博谣言识别的深度网络结构

将特征 f1′、f2 和 f3 拼接得到融合特征向量 X_f，即 $X_f=[f_1', f_2, f_3]=\{X_f^1, X_f^2, \cdots, X_f^i, \cdots, X_f^n\}$，$X_f$ 的向量维度 $n=|f1'|+|f2|+|f3|$。将 X_f 输入 BiLSTM 层得到正向输出与反向输出，分别为 $Lh=\{Lh_1, \cdots, Lh_i, \cdots, Lh_k\}$ 与 $Rh=\{Rh_1, \cdots, Rh_i, \cdots, Rh_k\}$。将正向和反向输出拼接得到 H，$H=[Lh, Rh]$，即 $|H|=2k$。定义全连接层的权值为 W_h，偏置为 b_h，输出 $H'=\{H_1', \cdots H_i' \cdots H_h'\}$。输出层用于判断是否为谣言，其权值为 w_o，偏置为 b_o，输出 $\hat{y}$，如公式(2)和公式(3)所示：

$$H'=W_h H+b_h \quad \text{公式 (2)}$$

$$\hat{y}=\sigma(W_o H'+b_o) \quad \text{公式 (3)}$$

5 实验及结果

5.1 实验设置

实验目标为微博中有关新冠疫情的谣言识别，设置了 4 个对照组分别检验不同情况下的实验结果以及笔者提出的模型的性能，具体各组实验如下：

(1)使用传统机器学习方法进行谣言识别。这类方法包括朴素贝叶斯(Naive Bayes, NB)、支持向量机(Support Vector Machine, SVM)、决策树(Decision Tree, DT)和集成学习(eXtreme Gradient Boosting, XGBoost)方法。

(2)使用有代表性的深度学习模型进行谣言检测。受数据量的限制，除了基本的深度学习模型卷积神经网络(Convolutional Neural Networks, CNN)以外，还需要特别的网络模型，这里选择迁移学习(Transform Learning, TL)模型和和生成对抗网络(Generative Adversarial Networks, GAN)模型。

(3)未使用 TF-IDF 进行文本增强的谣言识别。此时使用的是原始的预训练词向量 f1 与主题特征 f2、情感特征 f3 的融合，以比较文本增强的效果。

(4)使用 TF-IDF 增强后的谣言识别(即本文模型)，将增强后的词向量 f1′与主题特征 f2、情感特征 f3 融合。

5.2 实验过程

实验在 Deepin20 操作系统 + Python3.8 编程环境下进行，采用 Pytorch 深度学习框架构建谣言识别模型并进行模型训练。实验具体过程包括：

数据预处理：采用 jieba 分词器与哈尔滨工业大学的停用词表进行分词和去除停用词，并在分词词典中加入“疫情”“新冠”等突发公共卫生事件的背景词。采用 Word2Vec 模型的 CBOW 算法对文本进行向量化，将每个词语转化为一个维度为 300 的向量，每条样本中的词向量求和平均后作为初始的文本特征 f1。

利用 TF-IDF 算法计算词的重要程度权重，与初始的词向量进行点积，再进行求和平均。将增强后的特征向量作为网络输入，由于 TF-IDF 权重与词向量点积相乘后产生的数值极小，训练过程中有梯度消失，导致网络不收敛，故加入权重调节系数 C，可以有效克服该问题。最后利用公式(1)计算得到增强的文本特征向量 f1′；使用谣言文本语料训练 LDA 主题模型，之后计算每条谣言所属主题的概率分布，作为文本话题类别特征 f2；利用 Snownlp 对每条样本进行情感得分的计算，之后进行 z-score 归一化，将情感得分转换到 −1 到 1 之间，从而得到样本的情感特征 f3。

以文本特征向量 f1 为输入，分别利用传统机器学习模型 NB、SVM、DT、XGBoost 和深度学习模型 CNN、TL、GAN 构建分类器。其中，迁移学习 TL 利用了文献[16]的历史谣言数据集训练的 BiLSTM 网络来构建疫情谣言的分类器。对抗生成网络 GAN 利用人工生成了 3 000 条新的虚假信息，通过每补充 500 条生成数据分别测试，结果在有 1 500 条生成数据时效果好于其他情况，说明原始数据和生成数据的平衡对结果也有影响。

为了防止过拟合和提高模型鲁棒性，采用权重衰减方法对学习率进行衰减并在 LSTM 层加入 dropout 机制，模型的主要参数如表 4 所示：

表 4 实验参数设置

参数	参数值
Epoch	20
Dropout	0.5
Weight_decay	1e−5
Batch_size	64
分类层激活函数	sigmoid
学习率	0.01
BiLSTM 隐藏层层神经元个数	200
BiLSTM 隐藏层层数	5
DNN 全连接层神经元个数	100
权重调节系数 C	100
CBOW 词向量维度	300

以文本特征向量 f1 为输入，分别加入主题特征 f2 和情感特征 f3，利用 BiLSTM + DNN 分类模型构建分类器。

以增强后的文本特征向量 f1′为输入，分别加入主题特征 f2 和情感特征 f3，利用 BiLSTM + DNN 分类模型构建分类器，这样与第 3 组实验对比，反映词向量权重增强以后的结果。BiLSTM + DNN 网络训练过程的 loss 如图 7 所示：

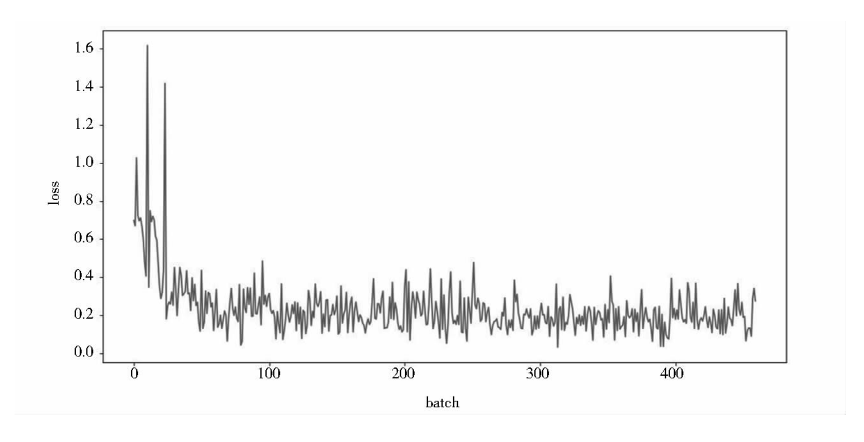


图 7　BiLSTM + DNN 网络 20 个 epoch 训练的 loss 曲线

5.3　实验结果与分析

在实验过程中,利用五折交叉验证方法,采用召回率、精确率和 F1 值衡量各模型性能,其中最重要的是召回率,即应该被识别的谣言有多大比例被识别出来。实验结果如表 5 所示:

表 5　突发公共卫生事件微博谣言识别结果

实验分组	Model	Recall	Precision	F1
第 1 组:与传统机器学习算法比较	NB	0.512 2	0.462 7	0.486 2
	SVM	0.547 8	0.587 0	0.566 7
	DT	0.663 0	0.749 2	0.703 4
	XGBoost	0.791 4	0.721 5	0.754 8
第 2 组:与深度学习模型的比较	CNN	0.738 2	0.636 5	0.683 6
	TL	0.828 6	0.778 1	0.802 6
	GAN	0.861 8	0.820 6	0.840 7
第 3 组:未进行关键词加强	f1	0.889 0	0.808 1	0.842 8
	f1 + f2	0.865 7	**0.980 5**	0.918 1
	f1 + f3	0.897 2	0.849 5	0.869 5
	f1 + f2 + f3	0.895 8	0.962 2	0.927 1
第 4 组:经过关键词权重的加强	f1′	0.932 8	0.858 6	0.891 3
	f1′ + f2	0.910 9	0.936 3	0.919 9
	f1′ + f3	0.932 8	0.804 0	0.858 4
	f1′ + f2 + f3(本文模型)	**0.960 2**	0.970 9	**0.965 4**

从表 5 可以看出,笔者所使用的方法和模型在识别任务中有较明显的优势。

从第 1 组实验可以发现,传统的机器学习模型识别结果召回率差别较大,集成学习算法效果较好,但也不到 80%。这一方面是受到训练数据集大小的限制;另一方面,这类模型较为简单,无法充分学习谣言的文本特征,使得模型对此类信息鉴别力较弱。

从第 2 组实验可以发现,深度学习的模型召回率有所提高,但其中 CNN 模型因为没有数据和特征优势,表现还不如传统机器学习模型。因为普通谣言数据和疫情谣言数据在文本特征、语言分布、领域对象等方面有较大差异,迁移学习方法的召回率也只有 83%。生成对抗学习结果最好,说明数据量的增加可以提高识别效果,但是原始数据只有 730 条,生成更多数据会打破与原始数据的平衡,效果反而下降。

从第 3 组实验可以明显看出,加入主题特征和情感特征后,模型的效果均有较大的提升,其中的话题类别特征(f2)对准确率和 f1 值的提升效果较为明显,情感特征 f3 对召回率的效果较好,而同时融合两种特征的模型取得的召回率、准确率和 F1 值的整体优势,说明加入的特征之间起到了一定的作用。

从第 4 组实验可以发现,经过对词向量的加权增强以后,再融合另外两组特征,在召回率和 F1 值均取得了最好的结果,精确率也在第二位。尤其是对谣言识别最重要的召回率,对比其他方法提升非常明显。同样,大部分实验的 F1 值也有明显提升,并且性能远

超原始词向量特征的模型。说明词向量加权对文本特征的强化作用,这样与其他特征组合后具有更好的互补作用,达到了较好的效果。

另外,本文实验还与一些已有的辟谣平台进行了比较,如中国互联网辟谣平台、微博辟谣、科学辟谣网等,但这些均为人工核实,缺少查证功能,无法起到早期发现谣言和遏制谣言的效果。而腾讯新闻的疫情谣言查证平台"较真网"则是根据用户输入与官方新闻的匹配进行谣言识别,缺少深度算法的支撑,多数本文实验能准确判别的谣言该网站还不能准确判别,这也说明了笔者提出的模型的实用性。

6 结语

突发公共卫生事件中的微博谣言识别对维系网络及社会稳定具有重要作用,在此背景下所做的谣言识别工作对将来应对更多的突发事件或公共卫生事件的谣言识别将带来积极的意义。由于独特的事件背景,传统的谣言识别方法效果有限。又因为受到数据量的限制,一般的深度学习模型也很难发挥作用。因此对文本特征的提取成为微博谣言识别的关键。笔者在语义词向量特征训练的基础上,新加入话题类别特征、情感特征以及关键词特征,设计了基于 TF-IDF 强化词向量表示的方法,合并另外两类特征以后,利用 BiLSTM + DNN 深度网络构建谣言识别模型,实验结果显示笔者提出的模型优于一些现有的相关模型,达到了较好的效果。进一步的工作可以通过加入更多的用户行为特征、时间序列特征、传播特征以及发布者身份特征等,目前来看这些特征数据集还不够完善,而且涉及太多过于复杂的特征处理过程,效果和效率还有待论证。另外,笔者所使用的词嵌入强化方法属于词粒度级别,在未来的研究中也可以探究不同粒度上或不同预处理的词嵌入强化表示的方法。

参考文献:

[1] 国务院政策网站. 突发公共卫生事件应急条例[EB/OL]. [2021-05-29]. http://www.gov.cn/zhengce/2020-12/26/content_5574586.htm.

[2] 张丽,毛浩然. 突发公共事件网络谣言治理研究综述[J]. 南京晓庄学院学报,2020,36(3):49-55,122.

[3] CARLOS C, MARCELO M, BARBARA P. Information credibility on twitter[C]//Proceedings of the 20th international conference on World Wide Web. New York: ACM, 2011: 675-684.

[4] YANG F, YU X, LIU Y, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on mining data semantics. New York: ACM, 2012:1-7.

[5] 贺刚,吕学强,李卓,等. 微博谣言识别研究[J]. 图书情报工作,2013,57(23):114-120.

[6] 夏松,林荣蓉,刘勘. 网络谣言敏感词库的构建研究-以新浪微博谣言为例[J]. 知识管理论坛,2019,4(5):267-275.

[7] 李钢,王聿达. 基于受众画像的新型耦合社交网络谣言传播模型研究[J]. 现代情报,2020,40(1):123-133,143.

[8] KWON S, CHA M, JUNG K, et al. Prominent features of rumor propagation in online social media[C]//2013 IEEE 13th international conference on data mining. Dallas: IEEE, 2013:1103-1108.

[9] MA J, GAO W, WEI Z, et al. Detect rumors using time series of social context information on Microblogging Websites[C]//ACM international conference on information and knowledge management. New York: ACM, 2015:1751-1754.

[10] 王志宏,过弋. 微博谣言事件自动检测研究[J]. 中文信息学报,2019,33(6):132-140.

[11] KOTTETI M, DONG X, QIAN L. Multiple time-series data analysis for rumor detection on social media[C]//2018 IEEE international conference on big data. Seattle: IEEE, 2018:4413-4419.

[12] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th international joint conference on artificial intelligence. New York: AAAI Press, 2016: 3818-3824.

[13] LI L, CAI G, CHEN N. A rumor events detection method based on deep bidirectional GRU neural network [C]//IEEE international conference on image. Piscataway: IEEE, 2018:755-759.

[14] 王星宇. 基于长短期记忆网络及深层特征的谣言识别系统[D]. 保定:河北大学,2020.

[15] 刘勘, 杜好宸. 基于深度迁移网络的 Twitter 谣言检测研究[J]. 数据分析与知识发现. 2019, 3 (10): 47-55.

[16] SIVA G, DEEPAK P, CHENG L, et al. Unsupervised fake news detection: a graph-based approach[C]// Proceedings of the 31st ACM conference on hypertext and social media. Florida: ACM, 2020: 75-83.

[17] MARRA F, GRAGNAIELLO D, COZZOLINO D, et al. Detection of GAN-generated fake images over social networks[C]//Proceedings of 2018 IEEE conference on multimedia information processing and retrieval. Piscataway: IEEE, 2018: 384-389.

[18] QIAN F, GONG C, SHARMA K, et al. Neural user response generator: fake news detection with collective user intelligence[C]// Proceedings of the 27th international joint conference on artificial intelligence. Stockholm: IJCAI, 2018: 3834-3840.

[19] 樊荣. 社会化媒体中突发公共事件语言传播特征及传播行为预测[D]. 大连:东北财经大学,2019.

[20] 曾子明,王婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例[J]. 情报学报,2019,38(1):89-96.

[21] 王林,王可,吴江. 社交媒体中突发公共卫生事件舆情传播与演变——以 2018 年疫苗事件为例[J]. 数据分析与知识发现,2019,3(4):42-52.

[22] 李丽华,韩思宁. 暴恐事件网络舆情传播机制及预防研究——英国典型案例的实证分析[J]. 情报杂志,2019,38(11):102－111,54.
[23] 曹娟,张勇东,李锦涛,等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报,2008(10):1780－1787.

作者贡献说明：
石锴文：获取数据、编程实验和撰写论文初稿；
刘勘：提出问题、设计思路和论文定稿。

Weibo Rumor Identification in Public Health Emergencies

Shi Kaiwen　Liu Kan
School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073

Abstract: [**Purpose/significance**] In public health emergencies such as the COVID-19 epidemic, a large number of statements about the epidemic have quickly been generated on social media on the Internet, including many rumors that endanger public mental health and affect the implementation of national policies. Detecting these remarks and identifying the rumors can enable the people to respond to public health emergencies correctly, and play a positive role in maintaining social stability and network governance. [**Method/process**] Firstly, the confirmed rumors during the epidemic were collected for in-depth analysis, and the main features of the rumor text were extracted, including context features, topic category features, sentiment level features, keyword features, etc.; then aiming at the problem that the text feature expression in the text classification model was relatively single, different models were used to vectorize the extracted rumor text features, and then a rumor recognition model based on multi-feature fusion was constructed. In the construction of this model, TF-IDF was used to strengthen the word vector, so that the word vector can merge the keyword feature information of the word granularity while capturing the context feature. Finally, this paper used the BiLSTM + DNN model to classify the fused feature vectors. [**Result/conclusion**] The experimental results show that features such as topic category and emotional level all contribute to the recognition of rumors, especially the fusion of the strengthened word vector and other features to significantly improve the recognition accuracy, recall rate, F1 measure, etc. The indicators all reached more than 90%, and the effect surpassed other rumor recognition models, indicating that the method constructed in this article can respond well to the task of rumor recognition in the context of public health emergencies.

Keywords: public health emergencies　rumor recognition　Weibo　multi-feature fusion